Statistical Science
2012, Vol. 27, No. 1, 31–50
DOI: 10.1214/11-STS355

Minimax and Adaptive Inference in Nonparametric Function Estimation

T. Tony Cai

Abstract. Since Stein's 1956 seminal paper, shrinkage has played a fundamental role in both parametric and nonparametric inference. This article discusses minimaxity and adaptive minimaxity in nonparametric function estimation. Three interrelated problems, function estimation under global integrated squared error, estimation under pointwise squared error, and nonparametric confidence intervals, are considered. Shrinkage is pivotal in the development of both the minimax theory and the adaptation theory.

While the three problems are closely connected and the minimax theories bear some similarities, the adaptation theories are strikingly different. For example, in a sharp contrast to adaptive point estimation, in many common settings there do not exist nonparametric confidence intervals that adapt to the unknown smoothness of the underlying function. A concise account of these theories is given. The connections as well as differences among these problems are discussed and illustrated through examples.

Key words and phrases: Adaptation, adaptive estimation, Bayes minimax, Besov ball, block thresholding, confidence interval, ellipsoid, information pooling, linear functional, linear minimaxity, minimax, nonparametric regression, oracle, separable rules, sequence model, shrinkage, thresholding, wavelet, white noise model.

1. INTRODUCTION

The multivariate normal mean model

$$x_i = \theta_i + \sigma z_i, \quad z_i \overset{\text{i.i.d.}}{\sim} N(0,1), \quad i = 1, \ldots, m, \tag{1}$$

occupies a central position in parametric inference. In his seminal paper, Stein (1956) showed that, when the dimension $m \geq 3$, the usual maximum likelihood estimator $Y = (y_i)$ of the normal mean is inadmissible under mean squared error

$$R(\hat\theta, \theta) = \frac{1}{m} \sum E(\hat\theta_i - \theta_i)^2, \tag{2}$$

and demonstrated that significant gain can be achieved by using shrinkage estimators. Since then shrinkage has become an indispensable technique in statistical inference, both in parametric and nonparametric settings.

This article considers minimaxity and adaptive minimaxity in nonparametric function estimation. Specifically, we discuss three interrelated problems: function estimation under global integrated squared error, estimation under pointwise squared error, and nonparametric confidence intervals. The goal is to give a concise account of important results in both the minimax theory and adaptation theory for each problem. The connections as well as differences among

T. Tony Cai is the Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA e-mail: tcai@wharton.upenn.edu.

these problems will be discussed and illustrated through examples. Shrinkage methods, including linear shrinkage, separable rules, thresholding and blockwise James–Stein procedures, figure prominently in the discussion.

A primary focus in nonparametric function estimation is the construction of adaptive procedures. The goal of adaptive inference is to construct a single procedure that achieves optimality simultaneously over a collection of parameter spaces. Informally an adaptive procedure automatically adjusts to the smoothness properties of the underlying function. A common way to evaluate such a procedure is to compare its maximum risk over each parameter space in the collection with the corresponding minimax risk.

As a step toward the goal of adaptive inference, one should first focus attention on the more concrete goal of developing a minimax theory over a given parameter space. This theory is now well developed particularly in the white noise with drift model:

$$dY(t) = f(t)\,dt + n^{-1/2}\,dW(t), \quad 0 \le t \le 1, \tag{3}$$

where $W(t)$ is a standard Brownian motion. This canonical white noise model is asymptotically equivalent to the conventional nonparametric regression where one observes (x_k, y_k) with

$$y_k = f(x_k) + z_k, \quad z_k \overset{\text{i.i.d.}}{\sim} N(0,1), \quad k = 1, \dots, n,$$

where $x_k = k/n$ in the fixed design case and $x_k \overset{\text{i.i.d.}}{\sim}$ Uniform$(0,1)$ in the case of random design. The parameter n in the white noise model (3) corresponds to the sample size in the regression model. See Brown and Low (1996a) and Brown et al. (2002). There is also a slightly less direct equivalence to density estimation and spectrum estimation. See Nussbaum (1996), Klemelä and Nussbaum (1999) and Brown et al. (2004).

Let $\{\beta_i(t), i \in \mathcal{I}\}$ be an orthonormal basis of $L^2[0,1]$ and let $y_i = \int \beta_i(t)\,dY_n(t)$ and $\theta_i = \int f(t)\beta_i(t)\,dt$. Then the white noise model (3) is equivalent to the following infinite-dimensional Gaussian sequence model

$$y_i = \theta_i + n^{-1/2} z_i, \quad z_i \overset{\text{i.i.d.}}{\sim} N(0,1), \quad i \in \mathcal{I}. \tag{4}$$

An estimator $\hat{\theta}$ of the mean sequence θ directly provides an estimator $\hat{f}(t) = \sum_{i \in \mathcal{I}} \hat{\theta}_i \beta_i(t)$ of the function f in the white noise model and vice versa. Hence, the function estimation model is closely related to the classical multivariate normal mean model (1). In these infinite-dimensional problems it is necessary to restrict the parameter set to be a compact subset of ℓ^2, the space of square summable sequences (or a compact subset of L^2, the space of square integrable functions, in the case of the white noise model). In contrast, the parameter set in the finite dimensional problem is typically all of $\mathbb{R}^m$.

Two of the most common ways of evaluating the performance of nonparametric function estimators are integrated squared error and pointwise squared error. Integrated squared error is used as a global measure of accuracy whereas pointwise squared error gives a local measure of loss. Minimax theory for both of these cases has been developed. We shall begin our discussion on minimax theory for estimation under integrated squared error. What follows will be elaborated in Section 2. Pinsker (1980) made a major breakthrough in nonparametric function estimation theory by giving a complete and explicit solution to the problem of minimax estimation over an ellipsoid under integrated squared error loss. Pinsker derived the minimax linear estimator and showed that the minimax risk is equal to the linear minimax risk asymptotically. Together these results yield the first precise evaluation of the asymptotic minimax risk in nonparametric function estimation. Donoho, Liu and McGibbon (1990) considered certain more general quadratically convex parameter spaces and showed that the linear minimax risk is within a small constant of the minimax risk. Furthermore, they also showed the limitations of linear procedures when the parameter space is not quadratically convex. Donoho and Johnstone (1998) studied minimax estimation over Besov balls which include cases that are not quadratically convex. Besov spaces are a very rich class of function spaces that are commonly used to model functions of inhomogeneous smoothness in functional analysis, statistics and signal processing. They also contain as special cases many traditional smoothness spaces such as Hölder and Sobolev spaces. The results of Donoho and Johnstone marked another major advance in the minimax estimation theory. In this setting it is shown that nonlinearity is essential for achieving minimaxity or even the minimax rate. Moreover, it is shown that the risk of the optimal coordinatewise thresholding rule is within a constant factor of the minimax risk.

The problem of estimating a function under pointwise squared error will be discussed in Section 4. This problem can be considered as a special case of estimating a linear functional. The minimax theory for estimating a linear functional over a convex parameter space has been well developed in Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994). In particular, the minimax difficulty of estimation is captured by a geometric quantity, the modulus of continuity, and the optimal linear shrinkage estimator is within a 1.25 multiple of the minimax risk. Cai and Low (2004a) extended this minimax theory to nonconvex parameter spaces. In this case, although the minimax rate of convergence is still determined by the modulus of continuity, optimal linear procedures can be arbitrarily far from being minimax and nonlinearity is necessary for minimax estimation.

The theory of adaptive estimation depends strongly on how risk is measured. When the performance is measured globally sharp adaptation can often be achieved. That is, one can attain the minimax risk over a collection of parameter spaces simultaneously. In particular, Efromovich and Pinsker (1984) constructed sharp adaptive estimators over a range of Sobolev spaces. Recent results on rate adaptive estimators focus on the more general Besov spaces. See, for example, Donoho and Johnstone (1995), Cai (1999), Johnstone and Silverman (2005) and Zhang (2005). In particular, Zhang (2005) developed general empirical Bayes methods which are asymptotically sharp minimax simultaneously over a wide collection of Besov balls. Adaptive estimation under the global loss will be discussed in Section 3. While separable rules are optimal for minimax estimation, they cannot be rate adaptive. Information pooling is a necessity for achieving adaptivity. Block thresholding provides a convenient and effective tool for information pooling. We discuss in detail block thresholding rules via the approach of ideal adaptation with an oracle. Through block thresholding, many shrinkage estimators developed in the normal decision theory can be used for nonparametric function estimation. In this sense block thresholding serves as a bridge between the classical theory and the modern function estimation theory.

Under pointwise risk it is often the case that sharp adaptation is not possible and a penalty, usually a logarithmic factor, must be paid for not knowing the smoothness. Important work in this area began with Lepski (1990) where attention focused on a collection of Lipschitz classes. Brown and Low (1996b) obtained similar results using a constrained risk inequality, Tsybakov (1998) investigated pointwise adaptation over Sobolev classes and Cai (2003) considered Besov spaces. Efromovich and Low (1994) studied estimation of linear functionals over a nested sequence of symmetric sets. A general adaptation theory for estimating linear functionals is given in Cai and Low (2005a). This theory gives a geometric characterization of the adaptation problem analogous to that given by Donoho (1994) for minimax theory. The adaptation theory describes exactly when rate adaptive estimators exist and when they do not exist the theory provides a general construction of estimators with the minimum adaptation cost.

In addition to point estimation, confidence sets also play a fundamental role in statistical inference. The construction of nonparametric confidence sets is an important and challenging problem. In Section 5 we consider nonparametric confidence sets with a particular focus on confidence intervals. Other confidence sets such as confidence balls and confidence bands have also been discussed in the literature. A minimax theory of confidence intervals for linear functionals was given in Donoho (1994) when the parameter space is assumed to be convex. Donoho (1994) constructed fixed length intervals centered at linear estimators which have length within a small constant factor of the minimax expected length. Cai and Low (2004b) extended the minimax theory for parameter spaces that are finite unions of convex sets. In this case it is shown that optimal confidence intervals centered at linear estimators can have expected length much larger than the minimax expected length. It is thus essential to center the confidence interval at a nonlinear estimator in order to achieve minimaxity over nonconvex parameter spaces.

An adaptation theory for confidence intervals was developed in Cai and Low (2004a). When attention is focused on adaptive inference there are some striking differences between adaptive estimation and adaptive confidence intervals. As mentioned earlier, sharp adaptation is often possible under integrated squared error and the cost of adaptation is typically a logarithmic factor under pointwise squared error. In contrast, in many common cases the cost of adaptation for confidence intervals is so high that adaptation becomes essentially impossible.

There is also a conspicuous difference between confidence intervals in parametric and nonparametric settings. To construct a confidence interval in parametric inference, a virtually universal technique is to first derive an optimal estimator of a parameter and then construct a confidence interval centered at this optimal estimator. It is often the case that such a method leads to an optimal confidence interval for the parameter. This is also a common practice in nonparametric function estimation. However, somewhat surprisingly, centering confidence intervals at optimally adaptive estimators in general yield suboptimal confidence procedures (Cai and Low, 2005c): Either the resulting interval has poor coverage probability or it is unnecessarily long.

The paper is organized as follows. We begin with minimax estimation under global integrated squared error loss. Section 2 focuses on the important results developed in Pinsker (1980), Donoho, Liu and McGibbon (1990) and Donoho and Johnstone (1998) on linear minimaxity, separable rules and minimaxity. Section 3 considers adaptive estimation under the global loss. The performance of separable rules is studied in the context of adaptive estimation. The results show that separable rules cannot be rate adaptive and information pooling is essential for adaptive estimation. We then discuss block thresholding rules using an oracle approach. Section 4 considers minimax and adaptive estimation under pointwise squared error loss and the construction of minimax and adaptive confidence intervals is treated in Section 5. The paper is concluded with discussions in Section 6.

2. LINEAR MINIMAXITY, SEPARABLE RULES AND MINIMAXITY

Minimax theory has been well developed in the Gaussian sequence model (and, equivalently, the white noise model). Two classes of estimators, namely, linear shrinkage rules and separable rules, figure prominently in the development of the theory. In this section we consider minimax estimation under global mean integrated squared error (MISE)

$$\begin{aligned} R(\hat f, f) &= E_f\|\hat f - f\|_2^2 \\ &= E_f\int_0^1 (\hat f(t) - f(t))^2\,dt \end{aligned} \tag{5}$$

for the function estimation model (3) and

$$R(\hat\theta,\theta) = E_\theta\|\hat\theta - \theta\|_2^2$$

for the sequence estimation model (4). Because of the isometry of the risks $R(\hat f, f) = R(\hat\theta,\theta)$ we shall focus on the sequence model (4) in this section. The performance of an estimator $\hat\theta$ over a parameter set $\mathcal{F}$ is measured by its maximum risk

$$R_n(\hat\theta,\mathcal{F}) = \sup_{\theta\in\mathcal{F}} E_\theta\|\hat\theta - \theta\|_2^2$$

and the benchmark is the minimax risk

$$R_n^*(\mathcal{F}) = \inf_{\hat\theta}\sup_{\theta\in\mathcal{F}} E_\theta\|\hat\theta - \theta\|_2^2.$$

When attention is restricted to linear procedures, we consider the linear minimax risk

$$R_n^L(\mathcal{F}) = \inf_{\hat\theta\ \text{linear}}\sup_{\theta\in\mathcal{F}} E_\theta\|\hat\theta - \theta\|_2^2.$$

In this section we give a concise account of some of the most important results in the minimax estimation theory without getting into too much technical detail. We refer interested readers to Iain Johnstone's monograph (Johnstone, 2002) for a detailed discussion on these and other related results.

2.1 Linear Minimaxity

Linear estimators and linear minimax risk occupy a special place in the development of nonparametric function estimation theory. Linear procedures are appealing because of their simplicity and linear minimax risk is easier to evaluate than the minimax risk. For example, for linear estimation over solid and orthosymmetric parameter spaces it suffices to focus on simple diagonal linear estimators of the form $\hat\theta_i = w_i y_i$ where w_i is a constant. Furthermore, in many settings the optimal linear procedure is asymptotically minimax or within a small constant of the minimax risk. See, for example, Pinsker (1980) and Donoho, Liu and McGibbon (1990). In this section we shall follow the historical development of the linear minimax theory by discussing the theory in the order of ellipsoids, quadratically convex classes and Besov classes.

Linear minimaxity over ellipsoids Pinsker (1980) considered minimax estimation over an ellipsoid

$$\mathcal{F} = \left\{\theta : \sum_{i=1}^{\infty} a_i^2\theta_i^2 \le M\right\}, \tag{6}$$

where $a_i \ge 0$ and $a_i \to \infty$. Since the ellipsoid $\mathcal{F}$ is symmetric, the linear minimax risk is attained by the optimal diagonal linear estimator of the form $\hat\theta(w) = (w_i y_i)$ where $w = (w_i) \in \ell^2$ with $0 \le w_i \le 1$ is a sequence of weights. That is,

$$R_n^L(\mathcal{F}) = \inf_{w}\sup_{\theta\in\mathcal{F}} E_\theta\|\hat\theta(w) - \theta\|_2^2. \tag{7}$$

The RHS of (7) is easy to evaluate. Note that

$$E_\theta\|\hat\theta(w) - \theta\|_2^2 = \sum_{i=1}^{\infty}(n^{-1}w_i^2 + (1-w_i)^2\theta_i^2).$$

Hence, the linear minimax risk

$$R_n^L(\mathcal{F}) = \inf_w \sup_{\theta\in\mathcal{F}} \sum_{i=1}^\infty (n^{-1}w_i^2 + (1-w_i)^2\theta_i^2) \tag{8}$$
$$= \sup_{\theta\in\mathcal{F}} \sum_{i=1}^\infty \frac{n^{-1}\theta_i^2}{n^{-1}+\theta_i^2}.$$

For any real number x, write $(x)_+$ for $\max(x,0)$. The Lagrange multiplier method shows that the maximum on the RHS of (8) is attained at $\theta_i^2 = n^{-1}(\mu/a_i - 1)_+$, where the parameter μ is determined by the constraint $\sum_{i=1}^\infty a_i^2\theta_i^2 = M$, which is equivalent to

$$n^{-1}\sum_{i=1}^\infty a_i(\mu - a_i)_+ = M.$$

The minimax linear estimator is given by $\hat{\theta}_{\text{l.minimax}} = (\hat{\theta}_i)$ with

$$\hat{\theta}_i = (1 - a_i/\mu)_+ y_i \tag{9}$$

and the linear minimax risk is

$$R_n^L(\mathcal{F}) = n^{-1}\sum_{i=1}^\infty (1 - a_i/\mu)_+. \tag{10}$$

A remarkable result of Pinsker (1980) is that for ellipsoidal $\mathcal{F}$ the linear minimax risk is asymptotically equal to the minimax risk, that is,

$$R_n^*(\mathcal{F}) = R_n^L(\mathcal{F})(1+o(1)).$$

Therefore, the minimax linear estimator $\hat{\theta}_{\text{l.minimax}}$ given in (9) is asymptotically minimax and the minimax risk is equal to the RHS of (10) asymptotically.

In the case of special interest where the parameter space is a Sobolev ball

$$\Theta_2^\alpha(M) = \left\{\theta : \sum_{k=1}^\infty (2\pi k)^{2\alpha}(\theta_{2k}^2 + \theta_{2k+1}^2) \le M\right\}$$

(which corresponds to a Sobolev ball in the function space under the usual trigonometric basis), the asymptotic minimax risk and the linear minimax risk can be evaluated explicitly as

$$R_n^*(\Theta_2^\alpha(M)) = R_n^L(\Theta_2^\alpha(M))(1+o(1)) \tag{11}$$
$$= \pi^{-2\alpha/(1+2\alpha)} M^{2/(1+2\alpha)} P_\alpha \cdot n^{-2\alpha/(1+2\alpha)}(1+o(1)),$$

where

$$P_\alpha = \left(\frac{\alpha}{1+\alpha}\right)^{2\alpha/(1+2\alpha)} (1+2\alpha)^{1/(1+2\alpha)}$$

is the Pinsker constant. This is the first exact evaluation of the asymptotic minimax risk in the nonparametric function estimation problem. See also Efromovich and Pinsker (1982) and Nussbaum (1985).

Pinsker's results represent a major contribution to nonparametric function estimation theory. Together they offer a complete and explicit solution to the problem of minimax estimation over ellipsoids.

Linear minimaxity over quadratically convex classes Donoho, Liu and MacGibbon (1990) considered certain more general quadratically convex parameter spaces. To discuss their results in more detail, we need first to introduce some terminology.

A parameter space $\mathcal{F}$ is called solid and orthosymmetric if $\theta = (\theta_1,\ldots,\theta_k,\ldots) \in \mathcal{F}$ implies that $\xi \in \mathcal{F}$ if $|\xi_i| \le |\theta_i|$ for all i. A set $\mathcal{F}$ is called quadratically convex if the set $\{(\theta_i^2)_{i=1}^\infty : \theta \in \mathcal{F}\}$ is convex. The quadratic convex hull of a set $\mathcal{F}$ is defined as

$$\text{Q.Hull}(\mathcal{F}) = \{(\theta_i)_{i=1}^\infty : (\theta_i^2)_{i=1}^\infty \in \text{Hull}(\mathcal{F}_+^2)\}, \tag{12}$$

where $\mathcal{F}_+^2 = \{(\theta_i^2)_{i=1}^\infty : (\theta_i)_{i=1}^\infty \in \mathcal{F}, \theta_i \ge 0\ \forall i\}$ and $\text{Hull}(\mathcal{F}_+^2)$ denotes the closed convex hull of the set $\mathcal{F}_+^2$.

Donoho, Liu and MacGibbon (1990) showed that for all solid orthosymmetric, compact and quadratically convex parameter spaces $\mathcal{F}$ the linear minimax risk is within a 1.25 factor of the minimax risk, that is,

$$R_n^L(\mathcal{F}) \le 1.25 R_n^*(\mathcal{F}). \tag{13}$$

Hence, the optimal linear procedure cannot be substantially improved by a nonlinear estimator. Donoho, Liu and MacGibbon (1990) proceeded by first solving an infinite-dimensional hyperrectangle problem where the parameter space $\mathcal{F}$ is of the form

$$\mathcal{F} = \{\theta : |\theta_i| \le \tau_i, i = 1,2,\ldots\} \tag{14}$$

with $\sum_i \tau_i^2 < \infty$. The traditional Hölder smoothness constraint in the function space corresponds to a hyperrectangle constraint in the sequence space with a suitably chosen (τ_i). See, for example, Meyer (1992). The problem of estimation over a hyperrectangle is solved by reducing it to coordinatewise one-dimensional bounded normal mean problems.

Consider estimating a bounded normal mean $\theta \in \mathbb{R}$ based on one observation $y \sim N(\theta,\sigma^2)$ with the prior knowledge that $|\theta| \le \tau$. It is easy to show that the minimax linear estimator of the bounded normal mean θ is

$$\delta^L(y) = \frac{\tau^2}{\tau^2+\sigma^2} y$$

and the minimax linear risk is

$$\rho^L(\tau,\sigma) \equiv \inf_{\delta\ \text{linear}} \sup_{|\theta|\le\tau} E_\theta(\delta(y)-\theta)^2 = \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}.$$

Denote the minimax risk for estimating the bounded normal mean θ by $\rho^*(\tau,\sigma)$. Let μ^* be the maximum value of the ratio of $\rho^L(\tau,\sigma)$ and $\rho^*(\tau,\sigma)$, that is,

$$\mu^* = \sup_{\tau,\sigma} \frac{\rho^L(\tau,\sigma)}{\rho^*(\tau,\sigma)}. \tag{15}$$

The constant μ^* is called the Ibragimov–Hasminskii constant. Ibragimov and Hasminskii (1984) studied the properties of the ratio $\rho^L(\tau,\sigma)/\rho^*(\tau,\sigma)$ and showed that the constant μ^* is finite. Donoho, Liu and MacGibbon (1990) proved that μ^* is in fact less than or equal to 1.25.

For estimation of θ over the hyperrectangle $\mathcal{F}$ given in (14) based on the sequence model (4), due to the independence of the observations y_i and the independent constraints on θ_i, it is not difficult to see that the minimax problem is separable. That is, the minimax (linear) estimator can be obtained through coordinatewise minimax (linear) estimation. Hence,

$$R_n^L(\mathcal{F}) = \sum_{i=1}^{\infty} \rho^L(\tau_i, n^{-1}) \quad \text{and}$$
$$R_n^*(\mathcal{F}) = \sum_{i=1}^{\infty} \rho^*(\tau_i, n^{-1})$$

and, consequently, for hyperrectangle $\mathcal{F}$,

$$R_n^L(\mathcal{F}) \le \mu^* R_n^*(\mathcal{F}) \le 1.25 R_n^*(\mathcal{F}). \tag{16}$$

A key step in solving the more general quadratically convex problem is to show that the difficulty for the linear estimators over the quadratically convex parameter space is in fact equal to the difficulty for the linear estimators of the hardest rectangular subproblem. Then (13) follows directly from (16).

In addition, Donoho, Liu and MacGibbon (1990) also showed that the linear minimax risk over a solid compact orthosymmetric set $\mathcal{F}$ is equal to that over the quadratic convex hull of $\mathcal{F}$,

$$R_n^L(\mathcal{F}) = R_n^L(\mathrm{Q.Hull}(\mathcal{F})). \tag{17}$$

This result indicates that although the optimal linear estimator is near minimax over quadratically convex parameter spaces, linear procedures have serious limitations when the parameter space $\mathcal{F}$ is not quadratically convex, especially when the quadratic convex hull of $\mathcal{F}$ is much larger than $\mathcal{F}$ itself. Such is the case in wavelet function estimation over certain Besov balls and in estimation of a sparse normal mean.

Linear minimaxity over Besov classes We now turn to wavelet estimation over Besov balls. It is more convenient to use double indices and write the sequence model (4) as

$$\begin{aligned} y_{j,k} &= \theta_{j,k} + n^{-1/2} z_{j,k}, \\ z_{j,k} &\overset{\text{i.i.d.}}{\sim} N(0,1), \quad (j,k) \in \mathcal{I}, \end{aligned} \tag{18}$$

where the index set $\mathcal{I} = \{(j,k): k = 1,\ldots,2^j, j = 0, 1, \ldots\}$. The Besov seminorm $\|\cdot\|_{b^\alpha_{p,q}}$ in the sequence space is then defined as

$$\|\theta\|_{b^\alpha_{p,q}} = \left(\sum_{j=0}^{\infty}\left(2^{js}\left(\sum_{k=1}^{2^j}|\theta_{j,k}|^p\right)^{1/p}\right)^q\right)^{1/q}, \tag{19}$$

where $s = \alpha + \frac{1}{2} - \frac{1}{p}$. We shall assume throughout the paper that $p, q, \alpha, s > 0$. The Besov ball $B^\alpha_{p,q}(M)$ is defined as a ball of radius M under this seminorm, that is,

$$B^\alpha_{p,q}(M) = \{\theta : \|\theta\|_{b^\alpha_{p,q}} \le M\}. \tag{20}$$

Besov spaces are a very rich class of function spaces and occur naturally in many areas of analysis. Besov spaces contain as special cases several traditional smoothness spaces such as Hölder and Sobolev spaces. For example, a Hölder space is a Besov space with $p = q = \infty$ and a Sobolev space is a Besov space with $p = q = 2$. Full details of Besov spaces are given, for example, in Triebel (1992) and DeVore and Lorentz (1993). See Meyer (1992) and Daubechies (1992) for wavelets and correspondence between function spaces and sequence spaces.

It is easy to verify that for $p \ge 2$ the Besov ball $B^\alpha_{p,q}(M)$ is quadratically convex and when $p < 2$,

$$\mathrm{Q.Hull}(B^\alpha_{p,q}(M)) = B^s_{2,q}(M), \tag{21}$$

where again $s = \alpha + \frac{1}{2} - \frac{1}{p}$. Besov spaces with $p < 2$ contain functions of a high degree of spatial inhomogeneity. See, for example, Triebel (1992), Meyer (1992) and DeVore and Lorentz (1993). Equations (21) and (17) together imply that for the Besov ball $B^\alpha_{p,q}(M)$ with $p < 2$,

$$\begin{aligned} R_n^L(B^\alpha_{p,q}(M)) &= R_n^L(\mathrm{Q.Hull}(B^\alpha_{p,q}(M))) \\ &= R_n^L(B^s_{2,q}(M)). \end{aligned} \tag{22}$$

In particular, for $p < 2$ the linear minimax risk over $B^\alpha_{p,q}(M)$ converges at the same rate as the minimax risk over $B^s_{2,q}(M)$. As we will see in Section 2.2, the minimax risk over $B^\alpha_{p,q}(M)$ converges at the rate of

$n^{-2\alpha/(1+2\alpha)}$ (Donoho and Johnstone, 1998). Since $s < \alpha$ for $p < 2$, $n^{-2s/(1+2s)} \gg n^{-2\alpha/(1+2\alpha)}$ and so the linear minimax risk over a Besov ball $B^\alpha_{p,q}(M)$ with $p < 2$ is substantially larger than the minimax risk. Therefore, the optimal linear estimator can be significantly outperformed by a nonlinear procedure. Intuitively, linear estimators do not perform well when the underlying functions are spatially inhomogeneous. In this case it is thus no longer desirable to restrict attention to the class of linear estimators.

Remark 1. It is interesting to note that a similar phenomenon also arises in the estimation of a quadratic functional. Cai and Low (2005b) showed that for estimating the quadratic functional $Q(\theta) = \sum_{i=1}^\infty \theta_i^2$ in the sequence model (4), the minimax quadratic risk over a solid orthosymmetric parameter space $\mathcal{F}$ equals the minimax quadratic risk over the quadratic convex hull of $\mathcal{F}$. Consequently, the optimal quadratic estimator of the quadratic functional $Q(\theta)$ is far from being minimax over a Besov ball $B^\alpha_{p,q}(M)$ with $p < 2$.

2.2 Separable Rules and Minimaxity

The shortcoming of linear procedures shows that nonlinearity is a necessity for achieving minimaxity over parameter spaces that are not quadratically convex, such as Besov balls $B^\alpha_{p,q}(M)$ with $p < 2$. Separable rules, which apply nonlinearity to individual coordinates separately, are a natural generalization of the linear shrinkage rules. Separable rules play a fundamental role in minimax estimation over parameter spaces that are not quadratically convex in a way similar to the role played by the linear estimators over the more conventional parametric spaces such as ellipsoids and hyperrectangles.

Under the sequence model (18), an estimator $\delta = (\delta_{j,k})$ is *separable* if for all $(j,k) \in \mathcal{I}$, $\delta_{j,k}$ depends solely on $y_{j,k}$, not on any other y's. We shall denote by $\mathcal{S}$ the collection of all separable rules. Well-known examples of separable rules include the traditional diagonal linear estimators, term-by-term thresholding estimators and Bayes estimators derived from independent priors. Separable rules are attractive because of their simplicity and intuitive appeal. More importantly, separable rules are minimax for a wide range of parameter spaces. In an important paper, Donoho and Johnstone (1998) pioneered the study of separable rules in minimax estimation over the Besov ball $B^\alpha_{p,q}(M)$ under the sequence model (18). Zhang (2005) further studied the class of separable rules in the context of sharp adaptation over the full scale of Besov balls using general empirical Bayes methods.

Donoho and Johnstone (1998) began by first solving the following minimax Bayes estimation problem. Suppose we observe $y = (y_{j,k})$ as in (18) with $\theta = (\theta_{j,k})$ itself a random vector satisfying a mean constraint

$$\|\tau\|_{b^\alpha_{p,q}} \le M,$$

where

$$\tau_{j,k} = (E|\theta_{j,k}|^{p\wedge q})^{1/(p\wedge q)}, \quad (j,k) \in \mathcal{I},$$

with $p \wedge q = \min(p,q)$. In other words, the "hard" constraint $\theta \in B^\alpha_{p,q}(M)$ in the original minimax problem is replaced by the "in mean" constraint $\tau \in B^\alpha_{p,q}(M)$ in the minimax Bayes problem. The minimax Bayes risk is defined as

$$R_n^B(B^\alpha_{p,q}(M)) = \inf_{\hat\theta} \sup_{\tau \in B^\alpha_{p,q}(M)} E\|\hat\theta - \theta\|_2^2.$$

Donoho and Johnstone (1998) showed that the minimax Bayes risk $R_n^B(B^\alpha_{p,q}(M))$ is attained by a separable rule $\hat\theta^* = (\hat\theta^*_{j,k})$ of the form

$$\hat\theta^*_{j,k} = \delta^*_j(y_{j,k}),$$

where $\delta^*_j(y_{j,k})$ is a scalar nonlinear function of $y_{j,k}$. Furthermore, when $\alpha + \frac{1}{2} > 1/(2 \wedge p \wedge q)$, the minimax Bayes risk is given by

$$\begin{aligned} &R_n^B(B^\alpha_{p,q}(M)) \\ &\quad = \gamma(Mn^{1/2}) M^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)} \\ &\qquad \cdot (1 + o(1)), \quad n \to \infty, \end{aligned} \tag{23}$$

where $\gamma(\cdot)$ is a continuous, positive, periodic function of $\log_2(Mn^{1/2})$. Moreover, when $p > q$, the minimax risk is asymptotically equal to the minimax Bayes risk,

$$R_n^*(B^\alpha_{p,q}(M)) = R_n^B(B^\alpha_{p,q}(M))(1 + o(1)),$$

and thus separable rules are minimax. Zhang (2005) further showed that the optimal separable rule is asymptotically minimax for general (p,q). In particular, these results showed that the minimax rate of convergence is n^{-r_*} where

$$r_* = \frac{\alpha}{\alpha + 1/2}. \tag{24}$$

That is,

$$\begin{aligned} 0 &< \varliminf_{n\to\infty} n^{r_*} R_n^*(B^\alpha_{p,q}(M)) \\ &\le \varlimsup_{n\to\infty} n^{r_*} R_n^*(B^\alpha_{p,q}(M)) < \infty. \end{aligned}$$

The linear minimax rate of convergence now follows immediately from (13), (17), (21) and (24). The linear minimax risk converges at the rate n^{-r_ℓ} where r_ℓ is given by

$$r_\ell = \frac{\alpha + (1/p_- - 1/p)}{\alpha + 1/2 + (1/p_- - 1/p)}, \qquad \text{where } p_- = \max(p,2).$$

It is clear that $r_\ell = r_*$ when $p \ge 2$ and $r_\ell < r_*$ when $p < 2$. Hence, nonlinear separable rules can outperform linear estimators at the level of convergence rates when $p < 2$.

2.3 Rate-Optimal Coordinatewise Thresholding Estimator

The separable minimax estimator that attains the minimax Bayes risk (23) is not available in closed form. Donoho and Johnstone (1998) showed that attention can be further restricted to a simpler coordinatewise thresholding estimator. It is shown that the optimal term-by-term thresholding estimator is within a small constant factor of the minimax risk. It was noted in Donoho and Johnstone (1998) that the constant factor is $\Lambda(p \wedge q) \le 1.6$ for $p \wedge q = 1$ using computational experiments and $\Lambda(p \wedge q) \le 2.2$ for $p \wedge q = 1$ for the essentially quadratically convex (and thus less important) case of $p \ge 2$. However, no specific rate optimal thresholding estimator is given in their paper.

We now present a rate-optimal coordinatewise thresholding estimator. Consider the sequence model (18). Let J_0 and J be integers satisfying, respectively, $M^{2/(1+2\alpha)} n^{1/(1+2\alpha)} \le 2^{J_0} < 2M^{2/(1+2\alpha)} n^{1/(1+2\alpha)}$ and $n \le 2^J < 2n$. For $j \ge J_0 + 1$, let

$$\lambda_j = \sqrt{2n^{-1}\log(2^{j-J_0})} \tag{25}$$

and let $\eta_\lambda(y) = \operatorname{sgn}(y)(|y| - \lambda)_+$ be the soft threshold function. We define the following thresholding estimator:

$$\hat\theta_{j,k} = \begin{cases} y_{j,k}, & \text{if } 1 \le j < J_0, \\ \eta_{\lambda_j}(y_{j,k}), & \text{if } J_0 \le j < J, \\ 0, & \text{if } j \ge J. \end{cases} \tag{26}$$

The estimator given in (26) is similar to the wavelet estimator given in Delyon and Juditsky (1996) for density estimation and nonparametric regression over $B^\alpha_{p,q}(M)$ under the Sobolev norm loss. It differs from the estimator in Delyon and Juditsky (1996) in the choice of the lower and upper resolution levels J_0 and J as well as in the choice of the thresholds λ_j. The following theorem can be shown using the same proof as given in Delyon and Juditsky (1996).

Theorem 1. *The separable estimator $\hat\theta$ given in (26) is within a constant factor of the minimax risk over the Besov ball $B^\alpha_{p,q}(M)$. That is,*

$$R_n(\hat\theta, B^\alpha_{p,q}(M)) \le C(\alpha,p,q) R_n^*(B^\alpha_{p,q}(M)),$$

where the constant $C(\alpha,p,q)$ depends only on α, p and q. In particular, the estimator is minimax rate-optimal,

$$\overline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} \sup_{\theta \in B^\alpha_{p,q}(M)} E\|\hat\theta - \theta\|_2^2 < \infty. \tag{27}$$

3. ADAPTIVE ESTIMATION THROUGH INFORMATION POOLING

Minimax risk provides a useful uniform benchmark for the comparison of estimators. However, the minimax estimators discussed in Section 2 require some explicit knowledge of the parameter space which is unknown in practice. A minimax estimator designed for a specific parameter space typically performs poorly over another parameter space. Recent work on nonparametric function estimation has focused attention on adaptive estimation, with the goal of constructing a single procedure which is near minimax simultaneously over a collection of parameter spaces. As mentioned in the Introduction, whether this goal can be accomplished depends strongly on how risk is measured. When the performance is measured by the global MISE risk sharp adaptation over Besov balls can be achieved. In fact, a large number of adaptive procedures have been developed in the literature. In this section we consider adaptive estimation under the MISE risk. For reasons of space, we do not give a comprehensive review of these adaptive estimators. We shall focus the discussion only on block thresholding which naturally connects shrinkage rules developed in the classical normal decision theory with nonparametric function estimation.

Because of the optimal performance of the separable rules in the minimax estimation setting, we begin in Section 3.1 by studying the adaptability of the separable rules. The results show that separable rules have their limitations; they cannot be rate adaptive, which implies that information pooling is the key to achieve adaptation. We then consider in Section 3.2 adaptive block thresholding estimators through ideal adaptation with oracle.

3.1 Adaptability of Separable Rules

As discussed in Section 2, Zhang (2005) showed that separable rules are asymptotically minimax over any given Besov ball $B^{\alpha}_{p,q}(M)$. Hence, from a minimax point of view there is little to gain by looking beyond the separable rules when the parameters (α, p, q) are fully specified. A natural question is whether separable rules can achieve the minimax rate of convergence simultaneously over a collection of Besov balls. To answer this question, we begin with a simple version of the adaptation problem by considering only two Besov balls. Let $B^{\alpha_1}_{p_1,q_1}(M_1)$ and $B^{\alpha_2}_{p_2,q_2}(M_2)$ be two Besov balls with $\alpha_1 \neq \alpha_2$. We call an estimator δ rate-adaptive over the two Besov balls if δ attains the minimax rate simultaneously over both of them, that is,

$$\begin{aligned} &\max_{i=1,2} \overline{\lim_{n\to\infty}}\, n^{2\alpha_i/(1+2\alpha_i)} \\ &\quad\cdot \sup_{\theta\in B^{\alpha_i}_{p_i,q_i}(M_i)} E\|\delta-\theta\|_2^2 < \infty. \end{aligned} \tag{28}$$

The question is: can (28) be achieved by a separable rule? To answer the question, Cai (2008) showed that separable rules are "inflexible": any rate-optimal separable rule over a Besov ball $B^{\alpha}_{p,q}(M)$ must have a "flat" rate of convergence everywhere in $B^{\alpha}_{p,q}(M)$. If a separable rule δ satisfies

$$\sup_{\theta\in B^{\alpha}_{p,q}(M)} E\|\delta-\theta\|_2^2 \le Cn^{-2\alpha/(1+2\alpha)}$$

for some constant $C > 0$, then for any given $\theta \in B^{\alpha}_{p,q}(M)$,

$$\begin{aligned} 0 &< \underline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} E\|\delta-\theta\|_2^2 \\ &\le \overline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} E\|\delta-\theta\|_2^2 < \infty. \end{aligned} \tag{29}$$

That is, δ must attain the exact same rate at every point $\theta \in B^{\alpha}_{p,q}(M)$. This is not the case for nonseparable rules. Indeed, there exist estimators that converge faster than the minimax rate at every point in $B^{\alpha}_{p,q}(M)$. See Brown, Low and Zhao (1997), Zhang (2005) and Cai (2008). As a direct consequence of the inflexibility of the separable rules, they are necessarily not rate-adaptive. That is, if $\alpha_1 \neq \alpha_2$, then

$$\begin{aligned} &\max_{i=1,2} \overline{\lim_{n\to\infty}}\, n^{2\alpha_i/(1+2\alpha_i)} \\ &\quad\cdot \inf_{\delta\in\mathcal{S}} \sup_{\theta\in B^{\alpha_i}_{p_i,q_i}(M_i)} E\|\delta-\theta\|_2^2 = \infty. \end{aligned} \tag{30}$$

The lack of adaptability of separable rules is closely connected to superefficiency in the classical univariate normal mean problem. It is well known that if an estimator of a univariate normal mean is superefficient at a point it must pay for the superefficiency by being subefficient in a neighborhood of that point. The Hodges estimator is an example of such estimators. See Le Cam (1953) and Brown and Low (1996b).

Under the sequence model (18), the minimax rate of convergence over the Besov ball $B^{\alpha}_{p,q}(M)$ is $n^{-2\alpha/(1+2\alpha)}$. We call an estimator δ *superefficient* at a fixed point $\theta \in B^{\alpha}_{p,q}(M)$ if

$$n^{2\alpha/(1+2\alpha)} E_\theta\|\delta-\theta\|_2^2 \to 0.$$

A heuristic proof of (29) sheds light on the cause of the lack of adaptability for separable rules. Let $\delta = (\delta_{j,k})$ be a minimax rate-optimal separable rule over $B^{\alpha}_{p,q}(M)$. Then individually each $\delta_{j,k}$ can be regarded as an estimator in a univariate normal mean problem. If δ is superefficient at some $\theta^* \in B^{\alpha}_{p,q}(M)$, then, as a univariate normal mean problem, many $\delta_{j,k}$ are superefficient at $\theta^*_{j,k}$ and, thus, each of these $\delta_{j,k}$ must be penalized in a subefficient neighborhood of $\theta^*_{j,k}$. There exists some $\theta' \in B^{\alpha}_{p,q}(M)$ with coordinates $\theta'_{j,k}$ in those subefficient neighborhoods of $\theta^*_{j,k}$. As a consequence of δ being superefficient at θ^*, δ is subefficient at θ' relative to the minimax risk over $B^{\alpha}_{p,q}(M)$. This contradicts the assumption that δ is rate-optimal uniformly over $B^{\alpha}_{p,q}(M)$. A rigorous argument can be found in Cai (2008). The main reason this phenomenon occurs is that separable rules estimate each coordinate $\theta_{j,k}$ based solely on an individual observation $y_{j,k}$. Estimation accuracy can be improved by pooling information on different coordinates to make more informative and accurate decisions.

Equation (30) shows that separable rules need to pay a price for adaptation. The minimum cost of adaptation for the separable rules is at least a logarithmic factor. Suppose $\alpha_1 > \alpha_2$. If a separable rule δ attains the minimax rate $n^{2\alpha_1/(1+2\alpha_1)}$ over $B^{\alpha_1}_{p_1,q_1}(M_1)$, then

$$\begin{aligned} &\underline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{2\alpha_2/(1+2\alpha_2)} \\ &\quad\cdot \sup_{\theta\in B^{\alpha_2}_{p_2,q_2}(M_2)} E\|\delta-\theta\|_2^2 > 0. \end{aligned} \tag{31}$$

This lower bound bears a strong similarity to the problem of adaptive estimation of a function at a point. See Section 4.

The lower bound (31) can indeed be attained by a separable rule. The well-known VisuShrink estimator of Donoho and Johnstone (1994) adaptively achieves within a logarithmic factor of the minimax risk. It is thus optimal among separable rules in the sense that it attains the lower bound on the adaptive convergence rate within this class of estimators.

To motivate the VisuShrink estimator, we begin with the classical multivariate normal mean model (1) and outline an oracle approach developed in Donoho and Johnstone (1994). Suppose we wish to estimate $\theta = (\theta_1, \ldots, \theta_m)$ based on the observations $x = (x_1, \ldots, x_m)$ in (1) under mean squared error (2).

In the discussion that follows, we focus on the separable rules. An ideal separable "estimator" $\hat{\theta}^{\text{ideal}}$ would estimate θ_i by x_i when $\theta_i^2 > \sigma^2$ and by 0 otherwise, that is, $\hat{\theta}_i^{\text{ideal}} = x_i I(\theta_i^2 > \sigma^2)$. This "estimator" achieves ideal trade-off between variance and squared bias for each coordinate and attains the ideal risk

$$R_{\text{DP.oracle}}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\theta_i^2 \wedge \sigma^2). \tag{32}$$

Since the "estimator" $\hat{\theta}^{\text{ideal}}$ requires the knowledge of the unknown θ, it is not a true statistical estimator. The ideal risk (32) is unattainable in practice, but it does provide a useful benchmark. To mimic the performance of the ideal "estimator" $\hat{\theta}^{\text{ideal}}$, Donoho and Johnstone (1994) proposed the soft threshold estimator

$$\hat{\theta}_i^* = \operatorname{sgn}(x_i)(|x_i| - \tau)_+, \tag{33}$$

with $\tau = \sigma\sqrt{2\log m}$, and showed the following Oracle Inequality:

$$\begin{aligned} &R(\hat{\theta}^*, \theta) \\ &\quad \le (2\log m + 1)[R_{\text{DP.oracle}}(\theta) + \sigma^2/m], \\ &\qquad\qquad \text{for all } \theta \in \mathbb{R}^m. \end{aligned} \tag{34}$$

Hence, the soft threshold estimator $\hat{\theta}^*$ comes within a logarithmic factor of the ideal risk for all $\theta \in \mathbb{R}^m$. Moreover, the factor $2\log m$ in the Oracle Inequality (34) is asymptotically sharp in the following sense:

$$\begin{aligned} &\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^m} \frac{E\|\hat{\theta} - \theta\|_2^2}{\sigma^2 + \sum_{i=1}^m \min(\theta_i^2, \sigma^2)} \\ &\quad = 2\log m(1 + o(1)), \quad m \to \infty. \end{aligned} \tag{35}$$

A similar result to (35) is given in Foster and George (1994) in the linear regression setting.

In the setting of the Gaussian sequence model (18), VisuShrink is defined as

$$\hat{\theta}_{j,k} = \begin{cases} \operatorname{sgn}(y_{j,k})(|y_{j,k}| - \sqrt{2n^{-1}\log n})_+, \\ \qquad \text{if } j < J, \\ 0, \quad \text{if } j \ge J, \end{cases} \tag{36}$$

where $J = \lfloor \log_2 n \rfloor$. The VisuShrink estimator adaptively achieves the rate of convergence $(\log n/n)^{2\alpha/(1+2\alpha)}$ over the Besov balls $B_{p,q}^{\alpha}(M)$ (Donoho et al., 1995). That is,

$$\sup_{\theta \in B_{p,q}^{\alpha}(M)} E\|\hat{\theta} - \theta\|_2^2 \le C\left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)}, \tag{37}$$

where $C > 0$ is a constant not depending on n. In light of the lower bound (31), VisuShrink is thus optimal within the class of separable rules.

3.2 Block Thresholding via Ideal Adaptation with Oracle

The results in Section 3.1 show that information pooling is a necessity for achieving full adaptation. Block thresholding, which estimates the coordinates in groups rather than individually, provides a convenient and effective tool for information pooling. Block thresholding increases estimation precision and achieves adaptivity by utilizing information about neighboring coordinates. The degree of adaptivity, however, depends on the choice of block size and threshold level.

We study block thresholding rules via the approach of ideal adaptation with an oracle. The main ideas of the oracle approach have been outlined at the end of Section 3.1 in developing the VisuShrink estimator. An oracle does not reveal the true estimand, but provides the ideal choice within a given class of estimators. The oracle "estimator" is typically not a true statistical estimator, as it may depend on the unknown parameter. It represents an ideal for a particular estimation method. The goal of ideal adaptation is to derive true statistical estimators which can essentially mimic the performance of an oracle.

The soft threshold estimator (33) estimates coordinates individually without using information about other coordinates. As we have shown in Section 3.1, such a separable rule is not optimal for adaptive estimation. We thus consider a more general class of estimators, the block projection (BP) estimators, which use information about neighboring coordinates by thresholding observations in groups. Simultaneous decisions are made to retain or discard all the coordinates within the same group.

We again begin with the finite-dimensional multivariate normal mean model (1). We wish to estimate the mean $\theta = (\theta_1, \ldots, \theta_m)$ based on the observations $x = (x_1, \ldots, x_m)$ in (1) under the mean squared error (2). Let $B_1, B_2, \ldots, B_N$ be a partition of the index set $\{1, \ldots, m\}$ with each B_i of size L (for convenience, we assume that the sample size m

is divisible by the block size L). Let $\mathcal{H}$ be a subset of the block indices $\{1,\dots,N\}$. A block projection estimator $\hat{\theta}(\mathcal{H})$ is defined as

$$\begin{aligned}\hat{\theta}_{B_j}(\mathcal{H}) = x_{B_j} \quad \text{if } j \in \mathcal{H} \quad \text{and} \\ \hat{\theta}_{B_j}(\mathcal{H}) = 0 \quad \text{if } j \notin \mathcal{H},\end{aligned} \tag{38}$$

where $x_{B_j} = (x_i)_{i\in B_j}$. The risk of $\hat{\theta}(\mathcal{H})$ is

$$\begin{aligned}&R(\hat{\theta}(\mathcal{H}),\theta) \\ &\quad = \frac{1}{m}\sum_{j=1}^{N}\{L\sigma^2 I(j\in\mathcal{H}) + \|\theta_{B_j}\|_2^2 I(j\notin\mathcal{H})\}.\end{aligned} \tag{39}$$

Ideally, one would like to choose $\mathcal{H}$ to consist of blocks j where $\|\theta_{B_j}\|_2^2 > L\sigma^2$. A BP oracle provides exactly this side information $\mathcal{H}_* = \mathcal{H}_*(\theta) = \{j : \|\theta_{B_j}\|_2^2 > L\sigma^2\}$, which yields the ideal block projection "estimator" $\hat{\theta}(\mathcal{H}_*)$ with $\hat{\theta}_{B_j}(\mathcal{H}_*) = x_{B_j} I(j \in \mathcal{H}_*)$ with the ideal risk

$$\begin{aligned}R_{\text{BP.oracle}}(\theta, L) &= \inf_{\mathcal{H}} \frac{1}{m} E\|\hat{\theta}(\mathcal{H}) - \theta\|_2^2 \\ &= \frac{1}{m}\sum_{j=1}^{N}(\|\theta_{B_j}\|_2^2 \wedge L\sigma^2).\end{aligned} \tag{40}$$

The ideal "estimator" $\hat{\theta}(\mathcal{H}_*)$ is not a true statistical estimator. A natural goal is to construct an estimator which can mimic the performance of the BP oracle.

Since Stein's 1956 seminar paper, many shrinkage estimators have been developed in the multivariate normal decision theory. Among them, the (positive part) James–Stein estimator is perhaps the best-known. Efron and Morris (1973) showed that the (positive part) James–Stein estimator does more than just demonstrate the inadequacy of the maximum likelihood estimator; it is a member of a class of good shrinkage rules, all of which may be useful in different estimation problems. Indeed, as we shall see below, blockwise James–Stein rules can essentially mimic the performance of the BP oracle when the threshold is properly chosen. For each block B_j let $S_j^2 = \sum_{i\in B_j} x_i^2$ and set

$$\hat{\theta}_{B_j}(L,\lambda) = \left(1 - \frac{\lambda L\sigma^2}{S_j^2}\right)_+ x_{B_j}. \tag{41}$$

Then the blockwise James–Stein estimator satisfies the following BP Oracle Inequality:

$$\begin{aligned}&R(\hat{\theta}(L,\lambda),\theta) \\ &\quad \le \lambda R_{\text{BP.oracle}}(\theta, L) + 4\sigma^2 \cdot P(\chi_L^2 > \lambda L),\end{aligned} \tag{42}$$

where χ_L^2 denotes a central chi-squared random variable with L degrees of freedom.

REMARK 2. When the block size $L = 1$, the estimator (41) becomes a coordinatewise thresholding estimator. It is easy to show that with the choice of $\lambda = 2\log m$ the BP Oracle Inequality (42) is equivalent to the Oracle Inequality (34) of Donoho and Johnstone (1994). The resulting estimator shares similar properties with the VisuShrink estimator. See Gao (1998).

REMARK 3. Another special choice of block size is $L = L_* = \log m$. The corresponding threshold is $\lambda = \lambda_* \equiv 4.50524$ (the solution of $\lambda - \log\lambda - 3 = 0$). The pair (L_*, λ_*) is chosen so that the corresponding estimator in the Gaussian sequence model is (near) optimal. See the discussion below. In this case the BP Oracle Inequality becomes

$$R(\hat{\theta}(L_*,\lambda_*),\theta) \le \lambda_* R_{\text{BP.oracle}}(\theta, L_*) + \frac{2\sigma^2}{m}. \tag{43}$$

Therefore, with block size $L_* = \log m$ and thresholding constant $\lambda_* = 4.50524$, the estimator comes essentially within a constant factor of 4.50524 of the ideal risk. Note that this blockwise James–Stein estimator is not minimax for a given block (since $\lambda_* > 2$), but it is close to being minimax and $\lambda_* = 4.50524$ is needed for the optimal performance in the infinite-dimensional Gaussian sequence model.

REMARK 4. Instead of the block projection estimators given in (38), one can also consider the more general block linear shrinkers: $\hat{\theta}_{B_j} = \gamma_j x_{B_j}, \gamma_j \in [0,1]$. In the case of block projection, $\gamma_j \in \{0,1\}$. An oracle would provide the ideal shrinkage factors $\gamma_j = \|\theta_{B_j}\|_2^2/(\|\theta_{B_j}\|_2^2 + L\sigma^2)$, and the ideal "estimator" has the risk

$$R_{\text{BLS.oracle}}(\theta, L) = \frac{1}{m}\sum_{j=1}^{N}\frac{\|\theta_{B_j}\|_2^2 L\sigma^2}{\|\theta_{B_j}\|_2^2 + L\sigma^2}.$$

The blockwise James–Stein estimator (41) also mimics the performance of the block linear shrinker oracle,

$$\begin{aligned}&R(\hat{\theta}(L,\lambda),\theta) \\ &\quad \le 2\lambda R_{\text{BLS.oracle}}(\theta, L) + 4\sigma^2 \cdot P(\chi_L^2 > \lambda L).\end{aligned} \tag{44}$$

We now return to the Gaussian sequence model (18) and consider the BlockJS procedure introduced in Cai (1999). Let $J = [\log_2 n]$. Divide each resolution level $1 \le j < J$ into nonoverlapping blocks of length $L = L_* = [\log n]$. (The coordinates in the first few resolution levels are grouped into a single block.)

Let b_i^j denote the set of indices of the coordinates in the ith block at level j, that is,

$$b_i^j = \{(j,k) : (i-1)L + 1 \le k \le iL\}.$$

Set $S_{j,i}^2 \equiv \sum_{k \in b_i^j} y_{j,k}^2$. We then apply the James–Stein shrinkage rule to each block b_i^j. For $(j,k) \in b_i^j$,

$$\hat{\theta}_{j,k}^* = \begin{cases} \left(1 - \dfrac{\lambda_* L n^{-1}}{S_{j,i}^2}\right)_+ y_{j,k}, \\ \qquad \text{for } (j,k) \in b_i^j, j < J, \\ 0, \quad \text{for } j \ge J, \end{cases} \tag{45}$$

where $\lambda_* \equiv 4.50524$ is the solution of $\lambda - \log\lambda - 3 = 0$. This threshold is derived based on the tail probability of a chi-squared distribution. See Cai (1999).

The BlockJS estimator (45) is adaptively within a constant factor of the minimax risk over all Besov balls $B_{p,q}^\alpha(M)$ for $p \ge 2$ and is within a logarithmic factor of the minimax risk over Besov balls $B_{p,q}^\alpha(M)$ for $p < 2$,

$$\begin{aligned} &\sup_{\theta \in B_{p,q}^\alpha(M)} E\|\hat{\theta}^* - \theta\|_2^2 \\ &\quad \le \begin{cases} C n^{-2\alpha/(1+2\alpha)} \\ \quad \text{for } p \ge 2 \\ C n^{-2\alpha/(1+2\alpha)} (\log n)^{(2/p-1)/(1+2\alpha)} \\ \quad \text{for } p < 2 \text{ and } \alpha p \ge 1. \end{cases} \end{aligned} \tag{46}$$

The block size and threshold level play important roles in the performance of a block thresholding estimator. The block size $L_* = \log n$ and threshold $\lambda_* = 4.50524$ are shown in Cai (1999) to be optimal in the sense that the resulting BlockJS estimator is both globally and locally adaptive. The extra logarithmic factor in the case of $p < 2$ is unavoidable for any block thresholding estimators with fixed block size and threshold.

Adaptation can be achieved through empirically selecting the block size and threshold at each resolution level by minimizing Stein's Unbiased Risk Estimate (Cai and Zhou, 2009). Let $y_{j\cdot} = (y_{j,1}, \ldots, y_{j,2^j})$. Since the positive part James–Stein estimator (41) is weakly differentiable, Stein's formula (Stein, 1981) for unbiased estimate of risk shows that

$$\begin{aligned} &\mathrm{SURE}(y_{j\cdot}, L, \lambda) \\ &\quad \equiv 2^j + \sum_i \frac{\lambda^2 L^2 - 2\lambda L(L-2)}{S_{(jb)}^2} \cdot I(S_{j,i}^2 > \lambda L) \\ &\qquad + (S_{j,i}^2 - 2L) \cdot I(S_{j,i}^2 \le \lambda L) \end{aligned}$$

is an unbiased estimate of the risk at level j. Choose the level-dependent block size L_j and threshold λ_j to be the minimizer of SURE:

$$(L_j, \lambda_j) = \operatorname*{arg\,min}_{L,\lambda} \mathrm{SURE}(y_{j\cdot}, L, \lambda).$$

The resulting estimator, called SureBlock, automatically adapts to the sparsity of the underlying sequence θ. In particular, the estimator is sharp adaptive over all Besov balls $B_{2,2}^\alpha(M)$ and simultaneously achieves within a factor of 1.25 of the minimax risk over Besov balls $B_{p,q}^\alpha(M)$ for all $p \ge 2$, $q \ge 2$. At the same time the SureBlock estimator achieves adaptively within a constant factor of the minimax risk over a wide collection of Besov balls $B_{p,q}^\alpha(M)$ in the "sparse case" $p < 2$. These properties are not shared simultaneously by other commonly used thresholding procedures such as VisuShrink (Donoho and Johnstone, 1994), SureShrink (Donoho and Johnstone, 1995) or BlockJS (Cai, 1999).

3.3 Discussion

The idea of block thresholding can be traced back to Efromovich (1985) in estimation using the trigonometric basis. A similar construction was used in Brown, Low and Zhao (1997) to produce superefficient estimators. In the context of wavelet estimation, global level-by-level thresholding was discussed in Donoho and Johnstone (1995) for regression and in Kerkyacharian, Picard and Tribouley (1996) for density estimation. Cavalier and Tsybakov (2002) and Cavalier et al. (2003) and Cai, Low and Zhao (2009) used weakly geometrically growing block size for sharp adaptation over ellipsoids. But these block thresholding methods are not local, they essentially adaptively mimic the performance of the ideal linear estimator. Because of the serious limitations of the linear procedures for estimating spatially inhomogeneous functions discussed at the end of Section 2.1, these estimators do not enjoy a high degree of spatial adaptivity. In particular, these estimators do not perform well over parameter spaces which are not quadratically convex such as Besov balls $B_{p,q}^\alpha(M)$ with $p < 2$.

Hall, Kerkyacharian and Picard (1998, 1999) introduced a local blockwise hard thresholding procedure for density estimation and nonparametric regression with a block size of the order $(\log n)^2$ where n is the sample size. Cai and Silverman (2001) considered overlapping block thresholding estimators. Block thresholding is a widely applicable technique. Cai and Low (2005b, 2006b) use block thresholding procedures for minimax as well as optimal adaptive estimation of a quadratic functional and Cai and Low (2006a) used a block thresholding method for the construction of adaptive confidence balls.

We have focused the discussion on blockwise James–Stein procedures because of their simplicity. In addition to the James–Stein rule, through block thresholding, many other shrinkage rules developed in the classical normal decision theory can be applied as well. For example, estimators of the forms

$$\hat{\theta} = [1 - \lambda_1 \sigma^2/(\lambda_2 + S^2)]_+ y \quad \text{or} \quad \hat{\theta} = [1 - c(S^2)]_+ y,$$

where $S^2 = \|y\|_2^2$ and $c(\cdot)$ is a suitably chosen function, can also be used. Besides block thresholding, the empirical Bayes method is another natural choice for information pooling and for constructing adaptive procedures. See Johnstone and Silverman (2005) and Zhang (2005). In particular, Zhang (2005) presented a class of general empirical Bayes estimators that are adaptively sharp minimax over a large collection of Besov balls. Other methods such as choosing a threshold by controlling the false discovery rate can also be used. See Abramovich et al. (2006).

4. MINIMAX AND ADAPTIVE ESTIMATION UNDER POINTWISE LOSS

So far the focus has been on the minimax and adaptive estimation under the global MISE risk (5). For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk measures such as (5) cannot wholly reflect the local performance of an estimator. The most commonly used measure of local accuracy is pointwise squared error loss. While the minimax theory under the pointwise loss is similar to that for the global loss, the adaptation theories for the two losses are significantly different. Under the local loss it is often the case that sharp adaptation is not possible and a penalty, usually a logarithmic factor, must be paid for not knowing the smoothness. Estimation under the pointwise risk (47) is a special case of estimating a linear functional $T(f)$. A general theory for estimating linear functionals has been developed in the literature. In this section we shall first focus on estimating a function under the pointwise risk and present a concise account of both the minimax and adaptation results. The related minimax and adaptation theory for estimating a general linear functional is discussed in Section 4.1.

We shall return to the white noise model (3) and consider estimation under pointwise squared error risk

$$R(\hat{f}, f; t_0) = E_f(\hat{f}(t_0) - f(t_0))^2, \tag{47}$$

where $t_0 \in (0,1)$ is any fixed point. For a given parameter space $\mathcal{F}$, the difficulty of the estimation problem is measured by the minimax risk

$$R_n^*(\mathcal{F}; t_0) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_f(\hat{f}(t_0) - f(t_0))^2. \tag{48}$$

Several methods have been developed to study the minimax estimation problem. These include modulus of continuity, metric entropy, information inequality, renormalization and constrained risk inequality. See, for example, Farrell (1972), Hasminskii (1979), Stone (1980), Ibragimov and Hasminskii (1984), Donoho and Liu (1991), Brown and Low (1991), Low (1992), Donoho and Low (1992) and Birgé and Massart (1995). For example, the minimax risk over any convex parameter space can be characterized, up to a small constant factor, in terms of the modulus of continuity. For estimation over the Besov balls, the minimax rate of convergence of the pointwise risk is derived in Cai (2003) using a constrained risk inequality. It is shown that the minimax risk satisfies

$$R_n^*(B_{p,q}^\alpha(M); t_0) \asymp n^{-2\nu/(1+2\nu)}, \tag{49}$$

where $\nu = \alpha - \frac{1}{p}$. Unlike the minimax rate of convergence under the global risk, the local minimax rate of convergence depends on the parameter p as well. Minimax rate optimal estimators can be constructed using wavelet thresholding.

The behavior of the estimators which are minimax rate optimal under the pointwise risk is quite different from that of rate optimal estimators under the global MISE risk. It is shown in Cai (2003) that if an estimator $\hat{f}$ attains the minimax rate of convergence over a Besov ball $B_{p,q}^\alpha(M)$, then it must attain the same "flat" rate at every f in the parameter space; superefficiency is not possible for rate optimal estimators. That is, if

$$\begin{aligned} &\overline{\lim_{n\to\infty}}\, n^{2\nu/(1+2\nu)} \\ &\quad \cdot \sup_{f \in B_{p,q}^\alpha(M)} E_f(\hat{f}(t_0) - f(t_0))^2 < \infty, \end{aligned} \tag{50}$$

then the estimator $\hat{f}$ must also satisfy

$$\underline{\lim_{n\to\infty}}\, n^{2\nu/(1+2\nu)} E_f(\hat{f}(t_0) - f(t_0))^2 > 0 \tag{51}$$

for any fixed $f \in B_{p,q}^\alpha(M)$. In contrast, under the global MISE risk, rate-optimal estimators over $B_{p,q}^\alpha(M)$ can achieve a much faster rate at some parameter points. Indeed, it is possible to have estimators which converge at a rate faster than the minimax rate at every fixed function in $B_{p,q}^\alpha(M)$; see Brown, Low and Zhao (1997), Zhang (2005) and Cai (2008).

Pioneering work on adaptive estimation under the pointwise risk began with Lepski (1990). This work focused on Lipschitz balls and showed that it is impossible to achieve complete adaptation for free when the smoothness parameter is unknown. One must pay a price for adaptation. Lepski (1990) and Brown and Low (1996b) showed that the cost of adaptation is at least a logarithmic factor even when the smoothness parameter is known to be one of two values. The case of the Sobolev balls was investigated by Tsybakov (1998). Cai (2003) considered adaptation over Besov balls.

The inflexibility of the minimax rate optimal estimators has direct consequence for adaptive estimation over Besov balls under the pointwise loss. Adaptation for free is only possible if the rates of convergence over the collection of the Besov balls are the same, that is, $\nu = \alpha - \frac{1}{p}$ is a fixed constant for all Besov balls in the collection. Otherwise, a penalty must be paid for adaptation, even over two Besov balls $B^{\alpha_i}_{p_i,q_i}(M_i)$, $i = 1, 2$. Let $\nu_i \equiv \alpha_i - 1/p_i$ for $i = 1, 2$ and suppose $\nu_1 > \nu_2 > 0$. If an estimator $\hat{f}$ attains a rate of n^ρ over $B^{\alpha_1}_{p_1,q_1}(M_1)$ with $\rho > 2\nu_2/(1 + 2\nu_2)$, in particular, if $\hat{f}$ is rate-optimal over $B^{\alpha_1}_{p_1,q_1}(M_1)$, then

$$\varliminf_{n\to\infty} \left(\frac{n}{\log n}\right)^{2\nu_2/(1+2\nu_2)} \cdot \sup_{f\in B^{\alpha_2}_{p_2,q_2}(M_2)} E_f(\hat{f}(t_0) - f(t_0))^2 > 0.$$

Therefore, the minimum cost for adaptation is at least a logarithmic factor. Furthermore, the rate $(n/\log n)^{2\nu/(1+2\nu)}$ can be adaptively attained, for example, by the VisuShrink estimator of Donoho and Johnstone (1994) and the BlockJS estimator discussed in Section 3.2. See Cai (2003).

Remark. We have focused on adaptation over different parameter spaces under a given loss. There is another type of adaptation problem which can be termed as loss adaptation: given a fixed parameter space, is it possible to construct an estimator that adapts to the loss function in the sense that the estimator is optimal both locally and globally? This problem was considered in Cai, Low and Zhao (2007). It was shown that it is impossible for any estimator to simultaneously attain the global minimax rate of convergence and the local minimax rate at every point when the global and local minimax rates are different. The minimum penalty for a global rate-optimal estimator is a logarithmic factor in terms of the maximum pointwise risk over $B^\alpha_{p,q}(M)$. The wavelet thresholding estimator with coefficients estimated by (26) is optimally loss adaptive in this sense.

4.1 Discussion on Estimation of Linear Functionals

The problem of estimating a function under the pointwise risk (47) is a special case of estimating a linear functional $T(f)$. For a given linear functional T and a parameter space $\mathcal{F}$ define the linear minimax risk $R^L_n(\mathcal{F}, T)$ and minimax risk $R^*_n(\mathcal{F}, T)$, respectively, by

$$R^L_n(\mathcal{F}, T) = \inf_{\hat{T}\ \text{linear}} \sup_{f\in\mathcal{F}} E_f(\hat{T} - T(f))^2 \quad \text{and}$$

$$R^*_n(\mathcal{F}, T) = \inf_{\hat{T}} \sup_{f\in\mathcal{F}} E_f(\hat{T} - T(f))^2.$$

The minimax theory for estimating a linear functional T over a convex parameter space has been well developed. See, for example, Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994). In particular, the properties of the minimax linear estimators can be described precisely and the linear minimax risk $R^L_n(\mathcal{F}, T)$ is within a small constant factor (≤1.25) of the minimax risk $R^*_n(\mathcal{F}, T)$, that is,

$$R^L_n(\mathcal{F}, T) \le \mu^* R^*_n(\mathcal{F}, T) \le 1.25 R^*_n(\mathcal{F}, T),$$

where μ^* is the Ibragimov–Hasminskii constant given in (15). A fundamental quantity which captures the difficulty of the estimation problem in this setting is the modulus of continuity

$$\begin{aligned}\omega(\varepsilon, \mathcal{F}) \\ = \sup\{|T(g) - T(f)| : \|g - f\|_2 \le \varepsilon, \\ f, g \in \mathcal{F}\}.\end{aligned} \tag{52}$$

For example, the linear minimax risk is given by

$$R^L_n(\mathcal{F}, T) = \sup_{\varepsilon>0} \frac{\omega^2(\varepsilon, \mathcal{F})}{4 + n\varepsilon^2} \tag{53}$$

and satisfies

$$\begin{aligned}\tfrac{1}{5}\omega^2(n^{-1/2}, \mathcal{F}) \le R^*_n(\mathcal{F}, T) \le R^L_n(\mathcal{F}, T) \\ \le \omega^2(n^{-1/2}, \mathcal{F}).\end{aligned}$$

See Ibragimov and Hasminskii (1984) and Donoho and Liu (1991).

In most common cases when estimating a linear functional over convex parameter spaces the modulus is Hölderian,

$$\omega(\varepsilon, \mathcal{F}) = C\varepsilon^{q(\mathcal{F})}(1 + o(1)). \tag{54}$$

In this case the exponent $q(\mathcal{F})$ determines the minimax rate of convergence. Hence, the rate of convergence is captured by the geometric quantity ω. Furthermore, Donoho and Liu (1991) showed that the modulus can be used to give a recipe for constructing the minimax linear estimator. A key step in this analysis is to show that the difficulty for linear estimators over a convex parameter space is in fact equal to the difficulty for linear estimators of the hardest one-dimensional subproblem. This problem is again closely connected to the problem of estimating a one-dimensional bounded normal mean discussed in Section 2.1. Cai and Low (2004a) extended the minimax theory for estimating linear functionals to nonconvex parameter spaces. It is shown that in this setting while the minimax rate of convergence is still determined by the modulus of continuity, the linear minimax risk can be arbitrarily far from the minimax risk. In fact, even if the parameter space is only a union of two convex sets, it is possible that the maximum risk of the best linear estimator does not even converge even though the minimax risk converges quickly. This shows that linear estimators have serious limitations when the parameter space is not convex.

The adaptation theory for estimating linear functionals is less well developed. As mentioned earlier, Lepski (1990) was the first to give examples which demonstrated that rate optimal adaptation over a collection of Lipschitz classes is not possible when estimating the function at a point. Efromovich and Low (1994) showed that this phenomena is true in general over a collection of nested symmetric sets. On the other hand, the goal of rate adaptive estimation of linear functionals can sometimes be realized. When the minimax rates over each parameter space are slower than any algebraic rate, Cai and Low (2003) have given examples of nested symmetric sets where sharp adaptive estimators can be constructed. In addition, when the parameter spaces are not symmetric, there are also examples where rate adaptive estimators can be constructed. See Efromovich (1997a, 1997b, 2000), Lepski and Levit (1998), Efromovich and Koltchinskii (2001) and Kang and Low (2002).

A general adaptation theory for estimating linear functionals is given in Cai and Low (2005a). This theory gives a geometric characterization of the adaptation problem analogous to that given by Donoho (1994) for minimax theory. This theory describes exactly when rate adaptive estimators exist, and when they do not exist the theory provides a general construction of estimators with minimum adaptation cost.

It is shown that two geometric quantities, a between class modulus of continuity and an ordered modulus of continuity, play a fundamental role in the adaptation theory. The between class modulus of continuity, defined by

$$\begin{aligned} &\omega_+(\varepsilon, \mathcal{F}_1, \mathcal{F}_2) \\ &\quad = \sup\{|T(g) - T(f)| : \|g - f\|_2 \le \varepsilon; \\ &\qquad\qquad f \in \mathcal{F}_1, g \in \mathcal{F}_2\}, \end{aligned} \tag{55}$$

captures the degree of adaptability over two convex parameter spaces in the same way that the usual modulus of continuity used by Donoho and Liu (1991) and Donoho (1994) captures the minimax difficulty of estimation over a single convex parameter space. The ordered modulus of continuity, given by

$$\begin{aligned} &\omega(\varepsilon, \mathcal{F}_1, \mathcal{F}_2) \\ &\quad = \sup\{T(g) - T(f) : \|g - f\|_2 \le \varepsilon; \\ &\qquad\qquad f \in \mathcal{F}_1, g \in \mathcal{F}_2\}, \end{aligned} \tag{56}$$

is instrumental in the construction of adaptive estimators with minimum adaptation cost.

The theory shows that there are three main cases in terms of the cost of adaptation. In the first case, the cost of adaptation is a logarithmic factor of n. This is the case for estimating a function at a point over Lipschitz balls. In the second case sharp adaptation is possible as in the examples considered in Lepski and Levit (1998) and Cai and Low (2003). This is also the case when estimating a convex or some other shape constrained function at a point. More dramatically, in the third case the cost of adaptation is much greater than in the first case. The cost of adaptation in this case is a power of n.

5. MINIMAX AND ADAPTIVE CONFIDENCE INTERVALS

The construction of confidence sets is an important part of statistical inference. As mentioned in the introduction, there are several types of nonparametric confidence sets including confidence intervals, confidence bands and confidence balls. For example, Li (1989), Beran and Dümbgen (1998), Genovese and Wasserman (2005), Cai and Low (2006a) and Robins and van der Vaart (2006) have constructed confidence balls with near optimal variable radius which also guarantee coverage probability. Adaptive confidence bands have been constructed in the special case of shape restricted functions. See Hengartner and Stark (1995) and Dümbgen (1998). See also Genovese and Wasserman (2008).

In this section we shall focus our discussion on pointwise confidence intervals for a function. Similar to estimation under the pointwise risk, this problem is a special case of confidence intervals for linear functionals. Both minimax theory and adaptation theory for confidence intervals of linear functionals have been developed. In this section we shall first discuss the general theory and then use confidence intervals for a function at a point as examples. Again, we will mainly use the Besov balls $B^\alpha_{p,q}(M)$ as the examples. The usual cases of Hölder balls and Sobolev balls follow by taking $p=q=\infty$ and $p=q=2$, respectively.

For any confidence interval there are two interrelated issues which need to be considered together, coverage probability and the expected length. A minimax theory for confidence intervals of linear functionals was given in Donoho (1994) for convex parameter spaces. In this setting the goal is to construct confidence intervals with a prespecified coverage probability which minimizes the expected length of the interval. Write $\mathcal{I}_{\gamma,\mathcal{F}}$ for the collection of all confidence intervals which cover the linear functional $T(f)$ with minimum coverage probability of $1-\gamma$ over the parameter space $\mathcal{F}$. Denote by

$$L(CI,\mathcal{F}) = \sup_{f\in\mathcal{F}} E_f(L(CI))$$

the maximum expected length of a confidence interval CI over $\mathcal{F}$ where $L(CI)$ is the length of CI. The benchmark is the minimax expected length of confidence intervals in $\mathcal{I}_{\gamma,\mathcal{F}}$,

$$L^*_\gamma(\mathcal{F}) = \inf_{CI\in\mathcal{I}_{\gamma,\mathcal{F}}} \sup_{f\in\mathcal{F}} E_f(L(CI)). \tag{57}$$

For convex $\mathcal{F}$, Donoho (1994) showed that the modulus of continuity defined in (52) determines the minimax expected length,

$$\begin{aligned} &2\omega(2z_\gamma n^{-1/2},\mathcal{F}) \\ &\quad\le L^*_\gamma(\mathcal{F}) \le 2\omega(2z_{\gamma/2} n^{-1/2},\mathcal{F}), \end{aligned} \tag{58}$$

where z_γ is the $100(1-\gamma)$th percentile of the standard normal distribution. Moreover, Donoho (1994) constructed fixed length intervals centered at linear estimators which have maximum length within a small constant factor of the minimax expected length $L^*_\gamma(\mathcal{F})$. Hence, from a minimax point of view there is relatively little to gain by centering the intervals on nonlinear estimators or using variable length intervals.

When the linear functional T is a point evaluation at $t_0\in(0,1)$, that is, $T(f)=f(t_0)$, and the parameter space is the Besov ball $B^\alpha_{p,q}(M)$, the modulus satisfies, with $\nu=\alpha-\frac{1}{p}$,

$$\omega(n^{-1/2},B^\alpha_{p,q}(M)) = Cn^{-\nu/(1+2\nu)}(1+o(1)).$$

Following the recipe given in Donoho (1994), one can construct a fixed length $1-\gamma$ level interval centered at a linear estimator with the length of order $n^{-\nu/(1+2\nu)}$.

The situation changes significantly when the parameter space is not convex. Cai and Low (2004a) developed a minimax theory for parameter spaces that are finite unions of convex parameter spaces. It is shown that in this case the optimal (variable length) confidence interval centered at linear estimators can have expected length much longer than the minimax expected length; it is thus essential to center the interval at nonlinear estimators in order to achieve optimality.

When attention is focused on adaptive inference there are some striking differences between adaptive confidence intervals and adaptive estimation. As we discussed in the earlier sections, adaptation for free is often possible under integrated squared error loss and the cost of adaptation is typically a logarithmic factor under pointwise squared error loss. For confidence intervals the cost of adaptation can be substantially more than that for estimation. In fact, in some common cases, the cost of adaptation is so high that adaptation becomes basically impossible. In these cases the maximum expected length of the confidence interval over any parameter space in the collection needs essentially to be equal to the maximum expected length over the whole collection in order for the confidence interval to have the desired coverage probability. See Low (1997).

An adaptation theory for confidence intervals was developed in Cai and Low (2004b). In light of the discussion on adaptive estimation given in Section 3, a natural goal for adaptive confidence intervals over a collection of parameter spaces $\{\mathcal{F}_i, i\in\mathcal{I}\}$ is to have a given coverage probability $1-\gamma$ over the union of the parameter spaces $\mathcal{F} = \bigcup_{i \in \mathcal{I}} \mathcal{F}_i$ and have the maximum expected length over each space within a constant factor of the corresponding minimax expected length, that is,

$$L(CI, \mathcal{F}_i) \le C_i L^*_\gamma(\mathcal{F}_i), \tag{59}$$

where C_i are constants. Unfortunately, in many common cases such adaptive confidence intervals do not exist even for two parameter spaces. Let $\{\mathcal{F}_1, \mathcal{F}_2\}$ be a pair of convex parameter spaces with nonempty intersection. Let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ and $0 < \gamma < \frac{1}{2}$. It is shown in Cai and Low (2004b) that for $i = 1, 2$

$$\begin{aligned} &\inf_{CI \in \mathcal{I}_{\gamma,\mathcal{F}}} L(CI, \mathcal{F}_i) \\ &\quad \ge \left(\frac{1}{2} - \gamma\right) \omega_+(z_\gamma n^{-1/2}, \mathcal{F}_i, \mathcal{F}), \end{aligned} \tag{60}$$

where the between class modulus ω_+ is defined in (55). The lower bound (60) can in fact be attained within a constant factor not depending on n. A general recipe, which relies on the ordered modulus $\omega(\varepsilon, \mathcal{F}_i, \mathcal{F}_j)$ defined in (56), is given in Cai and Low (2004b) for the construction of confidence intervals which attains the lower bound within a constant factor.

The lower bound (60), however, can be dramatically larger than the minimax expected length if the parameter space is prespecified. Such is the case for pointwise confidence intervals over Besov balls. Consider constructing a confidence interval for a function at a point $t_0 \in (0, 1)$ over two Besov balls based on the white noise model. In this case the linear functional $T(f) = f(t_0)$. Let $\mathcal{F}_i = B^{\alpha_i}_{p_i,q_i}(M_i)$ with $\nu_i \equiv \alpha_i - 1/p_i$ for $i = 1, 2$, $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ and suppose $\nu_1 > \nu_2 > 0$. Then standard calculations, as in, for example, Donoho and Liu (1987), show

$$\begin{aligned} \omega_+(\varepsilon, \mathcal{F}_i, \mathcal{F}) &= \omega(\varepsilon, \mathcal{F}) \\ &= C\varepsilon^{2\nu_2/(1+2\nu_2)}(1 + o(1)), \quad i = 1, 2. \end{aligned}$$

Thus, any $1 - \gamma$ level confidence intervals over both $B^{\alpha_1}_{p_1,q_1}(M_1)$ and $B^{\alpha_2}_{p_2,q_2}(M_2)$ must have the maximum expected length over $B^{\alpha_1}_{p_1,q_1}(M_1)$ satisfying

$$\begin{aligned} &L(CI, B^{\alpha_1}_{p_1,q_1}(M_1)) \\ &\quad \ge (\tfrac{1}{2} - \gamma)\omega_+(z_\gamma n^{-1/2}, B^{\alpha_1}_{p_1,q_1}(M_1), \mathcal{F}) \\ &\quad \asymp \omega(z_\gamma n^{-1/2}, \mathcal{F}) \\ &\quad \asymp n^{-\nu_2/(1+2\nu_2)}. \end{aligned} \tag{61}$$

In contrast, if it is known that $f \in B^{\alpha_1}_{p_1,q_1}(M_1)$, $1 - \gamma$ level confidence intervals can be constructed which satisfy $L(CI, B^{\alpha_1}_{p_1,q_1}(M_1)) \le Cn^{-\nu_1/(1+2\nu_1)} \ll Cn^{-\nu_2/(1+2\nu_2)}$. From (61), the rate of convergence of the maximum expected length of CI over $B^{\alpha_1}_{p_1,q_1}(M_1)$ is the same as that for the maximum expected length over $\mathcal{F}$. From this point of view the cost of adaptation is so high that adaptation is impossible.

It is also interesting to note an important difference between parametric confidence intervals and nonparametric intervals. In the parametric setting, a universal practice for the construction of a confidence interval is to first obtain an optimal estimator of a parameter and then construct a confidence interval for the parameter centered at this estimator. Such a method often leads to an optimal confidence interval for the parameter. That is, the confidence interval has a desired coverage probability and the length of the interval is the shortest. In nonparametric function estimation, it is also a common practice to center confidence intervals on optimally adaptive estimators. However, somewhat surprisingly, this in general leads to suboptimal confidence procedures (Cai and Low, 2005c). That is, either the confidence interval has poor coverage probability or it is unnecessarily long. It is instructive to consider an example.

Let us return to the problem of constructing a confidence interval for $f(t_0)$ over the two Besov balls $B^{\alpha_i}_{p_i,q_i}(M_i)$, $i = 1, 2$. Again let $\nu_i \equiv \alpha_i - 1/p_i$ for $i = 1, 2$ and suppose $\nu_1 > \nu_2 > 0$. Equation (61) shows that any confidence interval with coverage probability of at least $1 - \gamma$ over $B^{\alpha_2}_{p_2,q_2}(M_2)$ must have the maximum expected length of the order $n^{-\nu_2/(1+2\nu_2)}$ over both $B^{\alpha_1}_{p_1,q_1}(M_1)$ and $B^{\alpha_2}_{p_2,q_2}(M_2)$. This bound can easily be attained by using an optimal fixed length confidence interval. Now suppose $\hat{f}(t_0)$ is an adaptive estimator under the mean squared error. Then, in particular, $\hat{f}(t_0)$ has the maximum risk over $B^{\alpha_1}_{p_1,q_1}(M_1)$ converging at a rate n^{-r} where $r > \frac{2\beta_2}{1+2\beta_2}$. It follows from the results in Cai and Low (2005c) that any confidence interval CI centered at $\hat{f}(t_0)$ with coverage probability of at least $1 - \gamma$ over $B^{\alpha_2}_{p_2,q_2}(M_2)$ must satisfy for some constant $C > 0$

$$\begin{aligned} L(CI, B^{\alpha_2}_{p_2,q_2}(M_2)) &\ge C\left(\frac{\log n}{n}\right)^{\nu_2/(1+2\nu_2)} \\ &\gg n^{-\nu_2/(1+2\nu_2)}. \end{aligned} \tag{62}$$

Hence, confidence intervals centered at a mean squared error rate adaptive estimator must have a longer maximum expected length over $B^{\alpha_2}_{p_2,q_2}(M_2)$.

An interesting question is when adaptive confidence intervals exist? It can be seen easily by comparing the lower bound (60) with the bounds (58) for the minimax expected length that adaptive confidence intervals exist if and only if the moduli satisfy

$$\omega_+(\varepsilon, \mathcal{F}_i, \mathcal{F}) \asymp \omega(\varepsilon, \mathcal{F}_i), \quad i = 1, 2,$$

or, equivalently, $\omega(\varepsilon, \mathcal{F}_2) \le C_1 \omega(\varepsilon, \mathcal{F}_1) \le C_2 \omega_+(\varepsilon, \mathcal{F}_1, \mathcal{F}_2)$. In this case adaptive confidence intervals exist. These intervals have maximum expected length which can attain the same optimal rate of convergence as the minimax confidence interval over known $\mathcal{F}_i$. This is the case for certain shape restricted function spaces.

Consider constructing pointwise confidence intervals for monotonically decreasing Lipschitz functions. Again, in this case let $T(f) = f(t_0)$ with $0 < t_0 < 1$. Let $\mathcal{D}$ be the set of all decreasing functions on the unit interval and for $0 < \beta \le 1$ let

$$\begin{aligned} Lip^{\beta}(M) = \{ & f : [0,1] \to \mathbb{R}, \\ & |f(x) - f(y)| \le M|x - y|^{\beta}\}. \end{aligned} \tag{63}$$

Let $\mathcal{D}^{\beta}(M) = \mathcal{D} \cap Lip^{\beta}(M)$ be the collection of monotonically decreasing Lipschitz functions. Note that for $0 < \beta_2 < \beta_1 \le 1$, $\mathcal{D}^{\beta_1}(M) \subset \mathcal{D}^{\beta_2}(M)$. Let $\mathcal{F} = \bigcup_{0 \le \beta \le 1} \mathcal{D}^{\beta}(M)$. Then standard calculations yield

$$\begin{aligned} & \omega_+(\varepsilon, \mathcal{D}^{\beta}(M), \mathcal{F}) \\ & \quad = \omega(\varepsilon, \mathcal{D}^{\beta}(M)) \\ & \quad = (2\beta + 1)^{1/(2\beta+1)} M^{1/(2\beta+1)} \varepsilon^{2\beta/(2\beta+1)}. \end{aligned} \tag{64}$$

The adaptive confidence interval CI^* given in equation (34) of Cai and Low (2004b) has coverage probability of at least $1 - \gamma$ over $\mathcal{F}$ and satisfies for any $0 < \beta \le 1$

$$\begin{aligned} & L(CI^*, \mathcal{D}^{\beta}(M)) \\ & \quad \le 12(2\beta + 1)^{1/(2\beta+1)} M^{1/(2\beta+1)} z_{\gamma/2}^{2\beta/(2\beta+1)} \\ & \qquad \cdot n^{-\beta/(2\beta+1)}(1 + o(1)). \end{aligned} \tag{65}$$

Hence, the adaptive confidence interval CI^* simultaneously achieves with a constant factor of the minimax expected length over all $\mathcal{D}^{\beta}(M)$ with $0 < \beta \le 1$. Adaptive confidence intervals also exist for convex functions. See Cai and Low (2007).

6. CONCLUDING REMARKS

From linear estimators in Pinsker's solution to the ellipsoid problem to separable rules in Donoho and Johnstone's approach to minimax estimation over Besov balls to thresholding estimators such as blockwise James–Stein in adaptive wavelet estimation, shrinkage plays a pivotal role in both the minimax theory and the adaptation theory in nonparametric function estimation. In particular, block thresholding can be viewed as a bridge between the classical normal decision theory and nonparametric function estimation. Through block thresholding, many shrinkage estimators developed in the classical theory can be used for function estimation.

The three problems discussed in the paper are strongly connected. The minimax difficulty of estimation can be characterized by the modulus of continuity and the cost of adaptation is captured by the between class modulus. The linear minimaxity and minimaxity in these three problems are all linked to the one-dimensional bounded normal mean problem. In all three problems the performance of linear procedures is closely linked to the (quadratic) convexity of the parameter space. Linear shrinkage rules are near optimal when the parameter space is convex (quadratically convex in the case of global estimation), and linear procedures can be arbitrarily far from being minimax when the parameter space is not convex.

Although the minimax theories for the three problems are similar, the adaptation theories are remarkably different. Among the three problems, the adaptation results are most positive for estimation under the global MISE risk. In this case adaptation for free can be achieved. On the other hand, the results for adaptive confidence intervals are very pessimistic in general. The cost of adaptation is so high that adaptation over commonly used smoothness spaces is virtually impossible, although adaptation for free can be achieved over shape restricted spaces. These results indicate that, while the traditional smoothness constraint works well for estimation, it may not be a practical or correct formulation for the construction of adaptive nonparametric confidence intervals or bands. Alternative formulations are needed. Genovese and Wasserman (2008) is one step in this direction.

In this paper we have chosen to focus the discussion on the canonical white noise with drift model to avoid some of the nonessential technical complications. Parallel results hold for nonparametric regression and density estimation. We should emphasize that the discussion as well as the references given in this paper are by no means extensive. Interested readers are referred to Johnstone (2002) for further discussion and for a large number of additional references on estimation under global integrated squared error loss.

ACKNOWLEDGMENTS

We thank the referee for constructive comments which have helped to improve the presentation of the paper. Research supported in part by NSF FRG Grant DMS-0854973.

REFERENCES

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26** 1826–1856. MR1673280

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653

Brown, L. D. and Low, M. G. (1991). Information inequality bounds on the minimax risk (with an application to nonparametric regression). *Ann. Statist.* **19** 329–337. MR1091854

Brown, L. D. and Low, M. G. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958

Brown, L. D. and Low, M. G. (1996b). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965

Brown, L. D., Low, M. G. and Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25** 2607–2625. MR1604424

Brown, L. D., Cai, T. T., Low, M. G. and Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538

Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097. MR2102503

Cai, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924. MR1724035

Cai, T. T. (2003). Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statist. Sinica* **13** 881–902. MR1997178

Cai, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *J. Multivariate Anal.* **99** 421–436. MR2396972

Cai, T. and Low, M. (2007). Adaptive estimation and confidence intervals for convex functions. Technical report, Dept. Statistics, Univ. Pennsylvania.

Cai, T. T. and Low, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.* **31** 1140–1153. MR2001645

Cai, T. T. and Low, M. G. (2004a). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.* **32** 552–576. MR2060169

Cai, T. T. and Low, M. G. (2004b). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494

Cai, T. T. and Low, M. G. (2005a). On adaptive estimation of linear functionals. *Ann. Statist.* **33** 2311–2343. MR2211088

Cai, T. T. and Low, M. G. (2005b). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108

Cai, T. T. and Low, M. G. (2005c). Adaptive estimation of linear functionals under different performance measures. *Bernoulli* **11** 341–358. MR2132730

Cai, T. T. and Low, M. G. (2006a). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240

Cai, T. T. and Low, M. G. (2006b). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** 2298–2325. MR2291501

Cai, T. T., Low, M. G. and Zhao, L. H. (2007). Trade-offs between global and local risks in nonparametric function estimation. *Bernoulli* **13** 1–19. MR2307391

Cai, T. T., Low, M. G. and Zhao, L. H. (2009). Sharp adaptive estimation by a blockwise method. *J. Nonparametr. Stat.* **21** 839–850. MR2572586

Cai, T. T. and Silverman, B. W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā Ser. B* **63** 127–148. MR1895786

Cai, T. T. and Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569–595. MR2502643

Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123** 323–354. MR1918537

Cavalier, L., Golubev, Y., Lepski, O. and Tsybakov, A. (2003). Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems. *Theory Probab. Appl.* **48** 534–556.

Daubechies, I. (1992). *Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107

Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. MR1400080

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **303**. Springer, Berlin. MR1261635

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414

Donoho, D. L. and Liu, R. G. (1987). Geometrizing rates of convergence I. Technical Report 137, Dept. Statistics, Univ. California, Berkeley.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence. III. *Ann. Statist.* **19** 668–701. MR1105839

Donoho, D. L., Liu, R. C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. MR1062717

Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970. MR1165601

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369. MR1323344

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314. MR1611768

Efromovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Teory Probab. Appl.* **30** 557–661.

Efromovich, S. Y. and Pinsker, M. S. (1982). Estimation of square-integrable probability density of a random variable. *Probl. Inf. Transm.* **18** 19–38.

Efromovich, S. (1997a). Robust and efficient recovery of a signal passed through a filter and then contaminated by non-Gaussian noise. *IEEE Trans. Inform. Theory* **43** 1184–1191. MR1454946

Efromovich, S. (1997b). Density estimation for the case of supersmooth measurement error. *J. Amer. Statist. Assoc.* **92** 526–535. MR1467846

Efromovich, S. (2000). On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.* **9** 117–139. MR1780750

Efromovich, S. and Koltchinskii, V. (2001). On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** 2876–2894. MR1872847

Efromovich, S. and Low, M. G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Related Fields* **98** 261–275. MR1258989

Efromovich, S. Y. and Pinsker, M. S. (1984). Learning algorithm for nonparametric filtering. *Autom. Remote Control* **11** 1434–440.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180. MR0300360

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7** 469–488. MR1665666

Genovese, C. R. and Wasserman, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33** 698–729. MR2163157

Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875–905. MR2396818

Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26** 922–942. MR1635418

Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9** 33–49. MR1678880

Has'minskiĭ, R. Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. In *Contributions to Statistics* (J. Jureckova, ed.) 91–97. Reidel, Dordrecht. MR0561262

Hengartner, N. W. and Stark, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550. MR1332580

Ibragimov, I. A. and Hasminskii, R. Z. (1984). Nonparametric estimation of the values of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406. MR0739497

Johnstone, I. M. (2002). Function estimation and Gaussian sequence model. Unpublished manuscript.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560

Kang, Y.-G. and Low, M. G. (2002). Estimating monotone functions. *Statist. Probab. Lett.* **56** 361–367. MR1898714

Kerkyacharian, G., Picard, D. and Tribouley, K. (1996). L^p adaptive density estimation. *Bernoulli* **2** 229–247. MR1416864

Klemelä, J. and Nussbaum, M. (1999). Constructive asymptotic equivalence of density estimation and Gaussian white noise. Discussion Paper No. 53, Sonderforschungsbereich 373, Humboldt Univ., Berlin.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913

Lepski, O. V. and Levit, B. Y. (1998). Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.* **7** 123–156. MR1643256

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

Low, M. G. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems. *Ann. Statist.* **20** 545–554. MR1150360

Low, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412

Meyer, Y. (1992). *Wavelets and Operators. Cambridge Studies in Advanced Mathematics* **37**. Cambridge Univ. Press, Cambridge. MR1228209

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984–997. MR0803753

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16** 53–68. MR0624591

Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Pro-*

ceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I 197–206. Univ. California Press, Berkeley. MR0084922

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650

Triebel, H. (1992). *Theory of Function Spaces. II. Monographs in Mathematics* **84**. Birkhäuser, Basel. MR1163193

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469. MR1700239

Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33** 54–100. MR2157796